

# Palindromic complexity of trees

Srećko Brlek<sup>1</sup>, Nadia Lafrenière<sup>1</sup>, and Xavier Provençal<sup>2</sup>

<sup>1</sup>Université du Québec à Montréal, Montréal, Québec, Canada

<sup>2</sup>Université de Savoie, Chambéry, France

brlek.srecko@uqam.ca

lafreniere.nadia.2@courrier.uqam.ca

xavier.provençal@univ-savoie.fr

**Abstract.** We consider finite trees with edges labeled by letters on a finite alphabet  $\Sigma$ . Each pair of nodes defines a unique labeled path whose trace is a word of the free monoid  $\Sigma^*$ . The set of all such words defines the language of the tree. In this paper, we investigate the palindromic complexity of trees and provide hints for an upper bound on the number of distinct palindromes in the language of a tree.

**Keywords:** Words, Trees, Language, Palindromic complexity, Sidon sets

## 1 Introduction

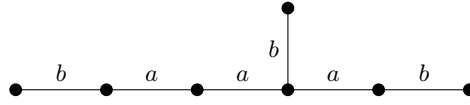
The palindromic language of a word has been extensively investigated recently, see for instance [1] and more recently [2,5]. In particular, Droubay, Justin and Pirillo [10] established the following property:

**Theorem 1** (Proposition 2 [10]) *A word  $w$  contains at most  $|w| + 1$  distinct palindromes.*

Several families of words have been studied for their total palindromic complexity, among which periodic words [4], fixed points of morphism [15] and Sturmian words [10].

Considering words as geometrical objects, we can extend some definitions. For example, the notion of palindrome appears in the study of multidimensional geometric structures, thus introducing a new characterization. Some known classes of words are often redefined as digital planes [3,16], and the adjacency graph of structures obtained by symmetries appeared more recently [9]. In the latter article, authors show that the obtained graph is a tree and its palindromes have been described by Domenjoud, Provençal and Vuillon [8]. The trees studied by Domenjoud and Vuillon [9] are obtained by iterated palindromic closure, just as Sturmian [7] and episturmian [10,13] words. It has also been shown [8] that the total number of distinct nonempty palindromes in these trees is equal to the number of edges in the trees. This property highlights the fact that these trees form a multidimensional generalization of Sturmian words.

A finite word is identified with a tree made of only one branch. Therefore, (undirected) trees appear as generalizations of words and it is natural to look forward to count the patterns occurring in it. Recent work by Crochemore et al. [6] showed that the maximum number of squares in a tree of size  $n$  is in  $\Theta(n^{4/3})$ . This is asymptotically bigger than in the case of words, for which the number of squares is known to be in  $\Theta(n)$  [12]. We discuss here the number of palindromes and show that, as for squares, the number of palindromes in trees is asymptotically bigger than in words. Figure 1, taken from [8], shows an example of a tree having more nonempty palindromes than edges, so that Theorem 1 does not apply to trees.



**Fig. 1.** A tree  $T$  with 6 edges and 7 nonempty palindromes, presented in [8].

Indeed, the number of nonempty factors in a tree is at most the ways of choosing a couple of edges  $(e_i, e_j)$ , and these factors correspond to the unique shortest path from  $e_i$  to  $e_j$ . Therefore, the number of nonempty palindromes in a tree cannot exceed the square of its number of edges. In this article, we exhibit a family of trees with a number of palindromes substantially larger than the bound given by Theorem 1. We give a value, up to a constant, for the maximal number of palindromes in trees having a particular language, and we conjecture that this value holds for any tree.

## 2 Preliminaries

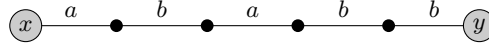
Let  $\Sigma$  be a finite alphabet,  $\Sigma^*$  be the set of finite words over  $\Sigma$ ,  $\varepsilon \in \Sigma^*$  be the empty word and  $\Sigma^+ = \Sigma^* \setminus \{\varepsilon\}$  be the set of nonempty words over  $\Sigma$ . We define the *language* of a word  $w$  by  $\mathcal{L}(w) = \{f \in \Sigma^* \mid w = pfs, p, s \in \Sigma^*\}$  and its elements are the *factors* of  $w$ . The *reverse* of  $w$  is defined by  $\tilde{w} = w_{|w|}w_{|w|-1} \dots w_2w_1$ , where  $w_i$  is the  $i$ -th letter of  $w$  and  $|w|$ , the length of the word. The number of occurrences of a given letter  $a$  in the word  $w$  is denoted  $|w|_a$ . A word  $w$  is a *palindrome* if  $w = \tilde{w}$ . The restriction of  $\mathcal{L}(w)$  to its palindromes is denoted  $\text{Pal}(w) = \{u \in \mathcal{L}(w) \mid u = \tilde{u}\}$ .

Some notions are issued from graph theory. We consider a *tree* to be an undirected, acyclic and connected graph. It is well known that the number of nodes in a tree is exactly one more than the number of edges. The *degree* of a node is given by the number of edges connected to it. A *leaf* is a node of degree 1. We consider a tree  $T$  whose edges are labeled by letters in  $\Sigma$ . Since in a tree there exists a unique simple path between any pair of nodes, the function  $p(x, y)$  that returns

the list of edges along the path from the node  $x$  to the node  $y$  is well defined, and so is the sequence  $\pi(x, y)$  of its labels. The word  $\pi(x, y)$  is called a *factor* of  $T$  and the set of all its factors, noted  $\mathcal{L}(T) = \{\pi(x, y) \mid x, y \in \text{Nodes}(T)\}$ , is called the *language* of  $T$ . As for words, we define the palindromic language of a tree  $T$  by  $\text{Pal}(T) = \{w \in \mathcal{L}(T) \mid w = \tilde{w}\}$ . Even though the *size* of a tree  $T$  is usually defined by its nodes, we define it here to be the number of its edges and denote it by  $|T|$ . This emphasizes the analogy with words, where the length is defined by the number of letters. Observe that, since a nonempty path is determined by its first and last edges, the size of the language of  $T$  is bounded by:

$$\mathcal{L}(T) \leq |T|^2 + 1. \quad (1)$$

Using the definitions above, we can associate a threadlike tree  $W$  to a pair of words  $\{w, \tilde{w}\}$ . We may assume that  $x$  and  $y$  are its extremal nodes (the leaves). Then,  $w = \pi(x, y)$  and  $\tilde{w} = \pi(y, x)$ . The size of  $W$  is equal to  $|w| = |\tilde{w}|$ . Analogously,  $\text{Pal}(W) = \text{Pal}(w) = \text{Pal}(\tilde{w})$ . The language of  $W$  corresponds to the union of the languages of  $w$  and of  $\tilde{w}$ . For example, Figure 2 shows the word  $ababb$  as a threadlike tree. Any factor of the tree is either a factor of  $\pi(x, y)$ , if the edges are read from left to right, or a factor of  $\pi(y, x)$ , otherwise.



**Fig. 2.** A threadlike tree represents a pair formed by a word and its reverse.

For a given word  $w$ , we denote by  $\Delta(w)$  its run-length-encoding, that is the sequence of constant block lengths. For example, for the French word “appelle”,  $\Delta(\text{appelle}) = 12121$ . As well, for the sequence of integers  $w = 11112111211211$ ,  $\Delta(w) = 4131212$ . Indeed, each letter of  $\Delta(w)$  represents the length of a block, while the length of  $\Delta(w)$  can be associated with the number of blocks in  $w$ . Given a fixed alphabet  $\Sigma$ , we define an infinite sequence of families of trees

$$\mathcal{T}_k = \{\text{tree } T \mid |\Delta(f)| \leq k \text{ for all } f \in \mathcal{L}(T)\}.$$

For any positive integer  $k$ , we count the maximum number of palindromes of any tree of  $\mathcal{T}_k$  according to its size. To do so, we define the function

$$\mathcal{P}_k(n) = \max_{T \in \mathcal{T}_k, |T| \leq n} |\text{Pal}(T)|.$$

This value is at least equal to  $n + 1$ . It is known [10] that each prefix  $p$  of a Sturmian word contains  $|p|$  nonempty palindromes. This implies that  $\mathcal{P}_\infty(n) \in \Omega(n)$ . On the other hand, equation (1) provides a trivial upper bound on the growth rate of  $\mathcal{P}_k(n)$  since it implies  $\mathcal{P}_\infty(n) \in \mathcal{O}(n^2)$ . We point out that  $\mathcal{P}_k(n)$  is an increasing function with respect to  $k$ . In the following sections we provide the asymptotic growth, in  $\Theta$ -notation, of  $\mathcal{P}_k(n)$ , for  $k \leq 4$ . Although we have not been able to prove the asymptotic growth for  $k \geq 5$ , we explain why we conjecture that  $\mathcal{P}_\infty(n) \in \Theta(\mathcal{P}_4(n))$  in section 5.

### 3 Trees of the family $\mathcal{T}_2$

First recall that, by definition, every nonempty factor of a tree  $T$  in  $\mathcal{T}_2$  has either one or two blocks of distinct letters. In other terms, up to a renaming of the letters, every factor in  $T$  is of the form  $a^*b^*$ . Therefore, any palindrome in  $T$  is on a single letter. From this, we can deduce a value for  $\mathcal{P}_2(n)$  :

**Proposition 2** *The maximal number of palindromes for the family  $\mathcal{T}_2$  is  $\mathcal{P}_2(n) = n + 1$ .*

*Proof.* The number of nonempty palindromes on a letter  $a$  is the length of the longest factor containing only  $a$ 's. Thus, the total number of palindromes is at most the number of edges in  $T$ , plus one (for the empty word). This leads directly to  $\mathcal{P}_2(n) \leq n + 1$ . On the other hand, a word of length  $n$  on a single-letter alphabet contains  $n + 1$  palindromes. This word is associated to threadlike tree in  $\mathcal{T}_1$ . Therefore,  $\mathcal{P}_2(n) = n + 1$ .  $\square$

### 4 Trees of the families $\mathcal{T}_3$ and $\mathcal{T}_4$

In this section, we show that  $\{\mathcal{P}_3(n), \mathcal{P}_4(n)\} \subseteq \Theta(n^{\frac{3}{2}})$ . To do so, we proceed in two steps. First, we present a construction that allows to build arbitrary large trees in  $\mathcal{T}_3$  such that the number of palindromes in their languages is large enough to show that  $\mathcal{P}_3(n) \in \Omega(n^{\frac{3}{2}})$ . Then, we show that, up to a constant, this construction is optimal for all trees of  $\mathcal{T}_3$  and  $\mathcal{T}_4$ .

#### 4.1 A lower bound for $\mathcal{P}_3(n)$ .

**Some elements from additive combinatorics.** An integer sequence is a *Sidon set* if the sums (equivalently, the differences) of all distinct pairs of its elements are distinct. There exists infinitely many of these sequences. For example, the powers of 2 are an infinite Sidon set. The maximal size of a Sidon set  $A \subseteq \{1, 2, \dots, n\}$  is only known up to a constant [14]. This bound is easily obtained since  $A$  being Sidon set, there are exactly  $\frac{|A|(|A|+1)}{2}$  sums of pairs of elements of  $A$  and all their sums are less or equal to  $2n$ . Thus,

$$\frac{|A|(|A|+1)}{2} \leq 2n$$

and  $|A| \leq 2\sqrt{n}$ . Erdős and Turán [11] showed that for any prime number  $p$ , the sequence

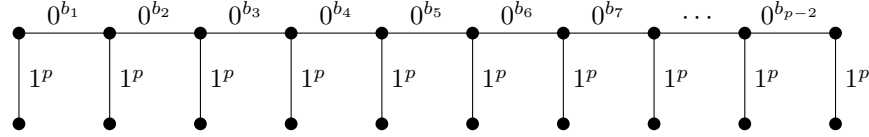
$$A_p = (2pk + (k^2 \bmod p))_{k=1,2,\dots,p-1}, \quad (2)$$

is a Sidon set. The reader should notice that, since there exists arbitrarily large prime numbers, there is no maximal size for sequences constructed in this way.

Moreover, the sequence  $A_p$  is, up to a constant, the densest possible. Indeed, the maximum value of any element of  $A_p$  is less than  $2p^2$  and  $|A_p| = p - 1$ . Since a Sidon set in  $\{1, 2, \dots, n\}$  is of size at most  $2\sqrt{n}$ , the density of  $A_p$  is  $\sqrt{8}$  (around 2.83) times smaller, for any large  $p$ .

**The hair comb construction.** Our goal is to describe a tree having a palindromic language of size substantially larger than the size of the tree. In this section, we build a tree  $\mathcal{C}_p \in \mathcal{T}_3$  for any prime  $p$  containing a number of palindromes in  $\Theta(|\mathcal{C}_p|^{\frac{3}{2}})$ .

For each prime number  $p$ , let  $B = (b_1, \dots, b_{p-2})$  be the sequence defined by  $b_i = a_{i+1} - a_i$ , where the values  $a_i$  are taken in the sequence  $A_p$  presented above, equation (2), and let  $\mathcal{C}_p$  be the tree constructed as follows :



**Proposition 3** *The sums of the terms in each contiguous subsequence of  $B$  are pairwise distinct.*

*Proof.* By contradiction, assume that there exists indexes  $k, l, m, n$  such that  $\sum_{i=k}^l b_i = \sum_{j=m}^n b_j$ . By definition of  $B$ ,

$$\sum_{i=k}^l b_i = \sum_{i=k}^l (a_i - a_{i-1}) = a_l - a_{k-1} \text{ and } \sum_{j=m}^n b_j = a_n - a_{m-1}.$$

This implies that  $a_l + a_{m-1} = a_n + a_{k-1}$ , which is impossible.  $\square$

**Lemma 4** *The number of palindromes in  $\mathcal{C}_p$  is in  $\Theta(p^3)$ .*

*Proof.* The nonempty palindromes of  $\mathcal{C}_p$  are of three different forms. Let  $c_0$  be the number of palindromes of the form  $0^+$ ,  $c_1$  be the number of palindromes of the form  $1^+$  and  $c_{101}$  be the number of palindromes of the form  $1^+0^+1^+$ . The number of palindromes of  $\mathcal{C}_p$  is clearly  $|\text{Pal}(\mathcal{C}_p)| = c_0 + c_1 + c_{101} + 1$ , where one is added for the empty word.

$$\begin{aligned} c_0 &= b_1 + b_2 + \dots + b_{p-2} = a_{p-1} - a_1 = 2p^2 - 4p, \\ c_1 &= p, \\ c_{101} &= |\{1^x 0^y 1^x \in \text{Pal}(\mathcal{C}_p)\}| \\ &= |\{x \mid 1 \leq x \leq p\}| \cdot |\{y \mid y = \sum_{i=k}^l b_i \text{ for } 1 \leq k \leq l \leq p-2\}| \\ &= \frac{1}{2}p(p-1)(p-2). \end{aligned}$$

The last equality comes from the fact that there are  $(p-1)(p-2)/2$  possible choices of pairs  $(k, l)$  and proposition 3 guarantees that each choice sums up to a different value. The asymptotic behavior of the number of palindromes is determined by the leading term  $p^3$ .  $\square$

**Lemma 5** *The number of edges in  $\mathcal{C}_p$  is in  $\Theta(p^2)$ .*

*Proof.* The number of edges labeled by 0 is  $b_1 + b_2 + \dots + b_{p-2} = 2p^2 - 4p$ . For those labeled with 1, there are exactly  $p-1$  sequences of edges labeled with 1's and they all have length  $p$ . The total number of edges is thus  $2p^2 - 4p + p(p-1) = 3p^2 - 5p$ .  $\square$

**Theorem 6**  $\mathcal{P}_3(n) \in \Omega(n^{\frac{3}{2}})$ .

*Proof.* Lemmas 4 and 5 implies that the number of palindromes in  $\mathcal{C}_p$  is in  $\Theta(|\mathcal{C}_p|^{\frac{3}{2}})$ . Since there are infinitely many trees of the form  $\mathcal{C}_p$  and since their size is not bounded, these trees provide a lower bound on the growth rate of  $\mathcal{P}_3(n)$ .  $\square$

#### 4.2 The value of $\mathcal{P}_4(n)$ is in $\Theta(n^{\frac{3}{2}})$ .

In this subsection, we show that the asymptotic value of  $\mathcal{P}_3(n)$  is reached by the hair comb construction, given above, and that it is the same value for  $\mathcal{P}_4(n)$ .

**Theorem 7**  $\mathcal{P}_4(n) \in \Theta(n^{\frac{3}{2}})$ .

Before giving a proof of this theorem, we need to explain some arguments. We first justify why we reduce any tree of  $\mathcal{T}_4$  to a tree in  $\mathcal{T}_3$ . Then, we present some properties of the latter trees in order to establish an upper bound on  $\mathcal{P}_4(n)$ .

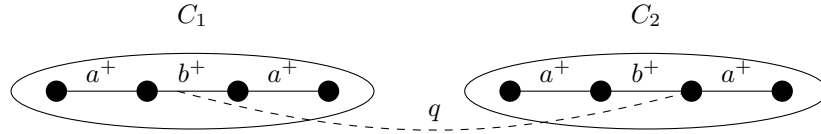
**Lemma 8** *For any  $T \in \mathcal{T}_4$ , there exists a tree  $S \in \mathcal{T}_3$  on a binary alphabet satisfying  $|S| \leq |T|$ , and with  $\frac{1}{|\Sigma|^2} |\text{Pal}(T)| - |T| \leq |\text{Pal}(S)| \leq |\text{Pal}(T)|$ .*

*Proof.* If there is in  $T$  no factor with three blocks starting and ending with the same letter, this means that all the palindromes are repetitions of a single letter. We then denote by  $a$  the letter on which the longest palindrome is constructed. It might not be unique, but it does not matter. Let  $S$  be the longest path labeled only with  $a$ 's. Then,  $|\text{Pal}(T)| \leq |\Sigma| |\text{Pal}(S)| \leq |\Sigma| |\text{Pal}(T)|$ . Otherwise, let  $a$  and  $b$  be letters of  $\Sigma$  and let  $(a, b)$  be a pair of letters for which  $|\mathcal{L}(T) \cap \text{Pal}(a^+b^+a^+)|$  is maximal. We define the set

$$E_S = \cup \{ p(u, v) \mid \pi(u, v) \in \text{Pal}(a^+b^+a^+) \}$$

and let  $S$  be the subgraph of  $T$  containing exactly the edges of  $E_S$  and the nodes connected to these edges. Then, there are three things to prove :

- $S$  is a tree: Since  $S$  is a subgraph of  $T$ , it cannot contain any cycle. We however need to prove that  $S$  is connected. To do so, assume that  $S$  has two connected components named  $C_1$  and  $C_2$ . Of course,  $\mathcal{L}(C_1) \subseteq a^*b^*a^*$  and  $C_1$  has at least one factor in  $a^+b^+a^+$ . The same holds for  $C_2$ . Since  $T$  is a tree, there is a unique path in  $T \setminus S$  connecting  $C_1$  and  $C_2$ . We call it  $q$ . There are paths in  $C_1$  and in  $C_2$  starting from an extremity of  $q$  and containing factors in  $b^+a^+$ . Thus, by stating that  $w$  is the trace of  $q$ ,  $T$  has a factor  $f \in a^+b^+a^*wa^*b^+a^+$ . By hypothesis,  $T \in \mathcal{T}_4$  so any factor of  $T$  contains at most four blocks. Then,  $f$  has to be in  $a^+b^+wb^+a^+$ , with  $w \in b^*$  and so  $q$  is a path in  $S$ . A contradiction.



- $S \in \mathcal{T}_3$  is on a binary alphabet: By construction,  $S$  contains only edges labeled by  $a$  or  $b$  and has no leaf connected to an edge labeled by  $b$ . This implies that if  $S$  contains a factor  $f \in a^+b^+a^+b^+$ ,  $f$  may be extended to  $f' \in a^+b^+a^+b^+a^+$ , which does not appear in  $T$ .
- $|\text{Pal}(S)| \geq \frac{1}{|\Sigma|^2} |\text{Pal}(T)| - |T|$ : We chose  $(a, b)$  to be the pair of letters for which the number of palindromes on an alphabet of size at least 2 was maximal. The number of palindromes on a single letter is at most  $|T|$ . Thus,

$$\frac{1}{|\Sigma|^2} |\text{Pal}(T)| - |T| \leq |\text{Pal}(S)| \leq |\text{Pal}(T)|.$$

□

**Lemma 9** *For any  $T \in \mathcal{T}_3$ ,  $T$  cannot contain both factors of  $0^+1^+0^+$  and of  $1^+0^+1^+$ .*

*Proof.* We proceed by contradiction. Assume that there exists in  $T$  four nodes  $u, v, x, y$  such that  $\pi(u, v) \in 0^+1^+0^+$  and  $\pi(x, y) \in 1^+0^+1^+$ . Since  $T$  is a tree, there exists a unique path between two nodes. In particular, there is a path from  $w \in \{u, v\}$  to  $w' \in \{x, y\}$  containing a factor of the form  $0^+1^+0^+\Sigma^*1^+$ , which contradicts the hypothesis that  $T \in \mathcal{T}_3$ . □

We now define the restriction  $\mathcal{R}_a(T)$  of a tree  $T$  to the letter  $a$  by keeping from  $T$  only the edges labeled by  $a$  and the nodes connected to them.

**Lemma 10** *Let  $T$  be in  $\mathcal{T}_3$ . There exists at least one letter  $a \in \Sigma$  such that  $\mathcal{R}_a(T)$  is connected.*

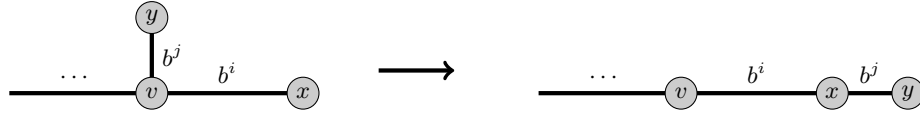
*Proof.* If  $T$  does not contain a factor on at least two letters that starts and ends with the same letter, that is of the form  $b^+a^+b^+$ , then  $\mathcal{R}_a(T)$  is connected for any letter  $a$ .

Otherwise, assume that a factor  $f \in b^+a^+b^+$  appears in  $T$ . Then,  $\mathcal{R}_a(T)$  must be connected. By contradiction, suppose there exists an edge labeled with  $a$  that is connected to the sequence of  $a$ 's in  $f$ , by a word  $w$  that contains another letter than  $a$ . Then, there exists a word of the form  $awa^+b^+$  in  $\mathcal{L}(T)$  and this contradicts the hypothesis that  $T \in \mathcal{T}_3$ .  $\square$

Given a node  $u$  in a tree, we say that  $u$  is a *splitting on the letter  $a$*  if  $\deg(u) \geq 3$  and there is at least two edges labeled with  $a$  connected to  $u$ .

**Lemma 11** *Let  $T$  be in  $\mathcal{T}_3$ . Then, there is a tree  $T'$  of size  $|T|$  such that  $\mathcal{L}(T) \subseteq \mathcal{L}(T')$  and there exists a letter  $a \in \Sigma$  such that any splitting of  $T'$  is on the letter  $a$ .*

*Proof.* If  $T$  is in  $\mathcal{T}_2$ , we apply the upcoming transformation to every branches. Otherwise, assume that a factor of the form  $b^+a^+c^+$  appears in  $T$  (note that  $b$  might be equal to  $c$ ). We allow splittings only on the letter  $a$ . Let  $v$  be a node of  $T$  that is a splitting on  $b \in \Sigma \setminus \{a\}$  (if it does not exist, then  $T' = T$ ). By the hypothesis on  $T$ , this means that there exists, starting from  $v$ , at least two paths labeled only with  $b$ 's leading to leaves  $x$  and  $y$ .



**Fig. 3.** The destruction of a splitting on the letter  $b$ .

We assume that  $|\pi(v, x)| \geq |\pi(v, y)|$ . Then, the words having  $\pi(v, y)$  as suffix are a subset of those for which  $\pi(v, x)$  is suffix. Therefore, the only case where  $\pi(v, y)$  may contribute to the language of  $T$  is when both the edges of  $\pi(v, x)$  and  $\pi(v, y)$  are used. The words of this form are composed only of  $b$ 's and are of length at most  $|\pi(v, x)| + |\pi(v, y)|$ . Moving the edges between  $s$  and  $y$  to the other extremity of  $x$ , we construct a tree for which the language contains  $\mathcal{L}(T)$  and having the same number of nodes. Finally, we can apply this procedure until the only remaining splittings are on the letter  $a$ . This leads to  $T'$ .  $\square$

We are now ready to prove the main theorem.

*Proof.* [Theorem 7:  $\mathcal{P}_4(n) \in \Theta(n^{\frac{3}{2}})$ .] Let  $T$  be in  $\mathcal{T}_4$ . By assumption, each factor of  $T$  contains at most four blocks of distinct letters.



1. Let  $S \in \mathcal{T}_3$  be such that  $|S| \leq |T|$ ,  $\mathcal{L}(S) \subseteq \{0, 1\}^*$  and  $\frac{|\text{Pal}(T)| - |T|}{|S|^2} \leq |\text{Pal}(S)| \leq |\text{Pal}(T)|$ . Using lemma 8, we know that this exists.

We know by lemma 9 that  $S$  may contain factors in  $1^+0^+1^+$ , but not in  $0^+1^+0^+$ .

2. By lemma 11, there exists a tree  $S'$  with  $|S'| = |S|$ , such that  $\mathcal{L}(S) \subseteq \mathcal{L}(S')$ , and with no splitting on the letter 1.

3. Finally, we count the palindromes in  $S'$ . The form of these palindromes is either  $0^+$ ,  $1^+$  or  $1^+0^+1^+$ . For the palindromes on a one-letter alphabet, their number is bounded by  $n$ , where  $n$  is the size of  $S'$ . We now focus on the number of palindromes of the form  $1^+0^+1^+$ . Call  $c_{101}$  this number. We show that  $c_{101} \leq 2n\sqrt{n}$ .

Since  $S'$  does not admit any splitting on the letter 1, each connected component of  $R_1(S')$  is a threadlike branch going from a leaf of  $S'$  to a node of  $R_0(S')$ . We name these connected components  $b_1, \dots, b_m$  and by lemma 10, we know that  $R_0(S')$  is connected.

Let  $b_i$  and  $b_j$  be two distinct branches of  $S'$ . By abuse of notation, we note  $\pi(b_i, b_j)$  the word defined by the unique path from  $b_i$  to  $b_j$ . Let  $l$  be such that  $\pi(b_i, b_j) = 0^l$  and suppose that  $|b_i| \leq |b_j|$ . Then, for any node  $u$  in  $b_i$ , there exists a unique node  $v$  in  $b_j$ , such that the word  $\pi(u, v) = 1^k 0^l 1^k$  is a palindrome. Moreover, if  $|b_i| < |b_j|$ , then there are nodes in  $b_j$  that cannot be paired to a node of  $b_i$  in order to form a palindrome. From this observation, a first upper bound is:

$$c_{101} \leq \sum_{1 \leq i < j \leq m} \min(|b_i|, |b_j|). \quad (3)$$

Another way to bound  $c_{101}$  is to count the palindromes of the form  $1^+0^+1^+$  according to the length of the block of 0's. For each length  $l$  from 1 to  $n$ , there might be more than one pair  $\{b_i, b_j\}$  that produces palindromes with central factor  $0^l$ . This provides a second upper bound:

$$c_{101} \leq \sum_{l=1}^n \max_{\substack{1 \leq i < j \leq m \\ \pi(b_i, b_j) = 0^l}} (\min(|b_i|, |b_j|)) \quad (4)$$

In order to obtain the desired bound on  $c_{101}$  we combine these two bounds. Let  $B' = \{i \mid |b_i| \geq \sqrt{n}\}$ . Since  $n$  is the size of  $S'$ , we have that  $|B'| \leq \sqrt{n}$  and that the average size of the branches  $b_i$  is such that  $i \in B'$  is bounded by  $n/|B'|$ . By applying the bound from (3) to the palindromes formed by two branches in  $B'$ , we obtain that the number of such palindromes is:

$$\sum_{\substack{1 \leq i < j \leq m \\ \{i, j\} \subseteq B'}} \min(|b_i|, |b_j|) \leq \frac{|B'|(|B'| - 1)}{2} \frac{n}{|B'|} \leq n\sqrt{n}. \quad (5)$$

Finally, it remains to count the number of palindromes that are defined by pairs of branches  $\{b_i, b_j\}$  such that  $i$  or  $j$  is not in  $B'$ . In such case, we always find that  $\min(|b_i|, |b_j|) < \sqrt{n}$ . The number of such palindromes is:

$$\sum_{l=1}^n \max_{\substack{1 \leq i < j \leq m \\ \pi(b_i, b_j) = 0^l \\ \{i, j\} \not\subset B'}} (\min(|b_i|, |b_j|)) < n\sqrt{n}. \quad (6)$$

Since each palindrome in  $S'$  is counted by equation (5) or (6), we obtain, summing both,  $c_{101} < 2n\sqrt{n} = 2|S'|^{\frac{3}{2}}$ . We deduce that, for any tree  $T$  in  $\mathcal{T}_4$ , the number of palindromes is bounded by

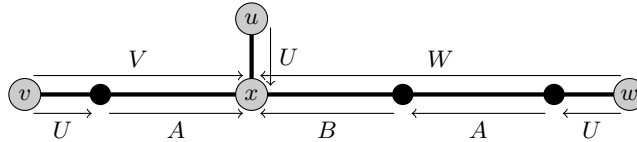
$$|\text{Pal}(T)| \leq |\Sigma|^2 |\text{Pal}(S)| + |T| < 2|\Sigma|^2 |S'|^{\frac{3}{2}} + |T| \leq 2|\Sigma|^2 |T|^{\frac{3}{2}} + |T|.$$

Using the fact that the alphabet is fixed (so its size is given by a constant), it is enough to prove that  $\mathcal{P}_4(n) \in \mathcal{O}(n^{\frac{3}{2}})$ . Combining this result with the one given in section 4.1, one may assert that both  $\mathcal{P}_3(n)$  and  $\mathcal{P}_4(n)$  are in  $\Theta(n^{\frac{3}{2}})$ .  $\square$

## 5 Hypotheses for the construction of trees with a lot of distinct palindromes

Let  $T$  be a tree that maximizes the number of palindromes for its size. It is likely that  $T$  contains triples of nodes  $(u, v, w)$  such that  $\pi(u, v)$ ,  $\pi(u, w)$  and  $\pi(v, w)$  are all palindromes. Suppose it is the case, and define  $T'$  as the restriction of  $T$  to the paths that join  $u$ ,  $v$  and  $w$ . We have that either  $T'$  is a threadlike tree, or  $T'$  has three leaves and a unique node of degree 3. The first case is of no interest here since it is equivalent to words, while the latter case implies a restrictive structure on the factors  $\pi(u, v)$ ,  $\pi(u, w)$  and  $\pi(v, w)$ . We now focus on the second case and call  $x$  the unique node of  $T'$  with degree 3.

Let  $U = \pi(u, x)$ ,  $V = \pi(v, x)$ ,  $W = \pi(w, x)$  and, without loss of generality, suppose that  $|U| \leq |V| \leq |W|$ . Then, as shown in Figure 4,  $U\tilde{V}$ ,  $U\tilde{W}$  and  $V\tilde{W}$  are all palindromes.



**Fig. 4.** The structure of the tree  $T'$ . The palindromicity of  $U\tilde{V}$ ,  $U\tilde{W}$  and  $V\tilde{W}$  forces that  $V$  starts with  $U$  while  $W$  starts with both factors  $U$  and  $V$ .

Let  $A$  be the suffix of length  $|V| - |U|$  of  $V$ . Since, by hypothesis,  $U\tilde{V}$  is a palindrome,  $V = UA$  and  $A$  is a palindrome. Similarly, let  $B$  be the suffix of length  $|W| - |V|$  of  $W$ . This implies that  $W = VB = UAB$  and both  $B$  and  $AB$  are palindromes. Using a well-known lemma from Lothaire [17], we prove that  $AB$  is periodic.

**Lemma 12** (Proposition 1.3.2 in [17]) *Two words commute if and only if they are powers of the same word.*

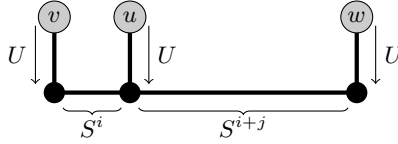
The next proposition states that the word  $ABA$  is periodic and that its period is at most the gcd of the difference of length of the three paths between  $u$ ,  $v$  and  $w$ . More formally, let

$$p = \gcd(|\pi(u, w)| - |\pi(u, v)|, |\pi(v, w)| - |\pi(u, v)|, |\pi(v, w)| - |\pi(u, w)|).$$

**Proposition 13** *There exists a word  $S$  and two integers  $i, j$  such that  $|S|$  divides  $p$  and  $A = S^i$  and  $B = S^j$ .*

*Proof.* Since  $A$ ,  $B$  and  $AB$  are palindromes,  $AB = \widetilde{AB} = \tilde{B}\tilde{A} = BA$ . Thus, by lemma 12, there exists a word  $S$  such that  $A = S^i$  and  $B = S^j$ . This implies that  $|S|$  divides  $\gcd(|A|, |B|)$  and, by construction,  $\gcd(|A|, |B|) = p$ .  $\square$

From the above proposition, we deduce that a triple of nonaligned nodes with any path from a node to another being a palindrome forces a local structure isomorphic to that of the hair comb tree, as illustrated in Figure 5.



**Fig. 5.** A triple of nodes with palindromes between each pair of them is isomorphic to a part of a hair comb.

In a more general way, suppose that a tree contains  $m$  leaves  $(u_i)_{1 \leq i \leq m}$ , and that each  $\pi(u_i, u_j)$  is a palindrome. Let  $T'$  be the restriction of this tree to the paths that connect these leaves and, for each  $i$ , let  $v_i$  be the first node of degree higher than 2 accessible from the leaf  $u_i$  in  $T'$ . By applying the above proposition to each triplet  $(u_i, u_j, u_k)$ , for all  $i \neq j$ , the word  $\pi(u_i, u_j)$  is of the form

$$\pi(u_i, u_j) = US^+\tilde{U},$$

where  $|U| = \min_i(\pi(u_i, v_i))$  and  $|S|$  divides  $\gcd_{i \neq j, k \neq l} (|\pi(u_i, u_j)| - |\pi(u_k, u_l)|)$ .

Moreover, in order to maximize the number of palindromes relatively to the size of the tree, we can choose  $S$  to be a single letter. This is indeed possible since the only condition on the length of  $S$  is that it divides all the differences of lengths between any palindromic path from a leaf to another.

This gives a tree analogous to those presented in section 4.1,  $\mathcal{C}_p$ , and for which we have established that  $|\text{Pal}(\mathcal{C}_p)| \in \Theta(|\mathcal{C}_p|^{\frac{3}{2}})$ . Therefore, we conjecture that  $\mathcal{P}_\infty(n) \in \Theta(n^{\frac{3}{2}})$ .

## References

1. Allouche, J.P., Baake, M., Cassaigne, J., Damanik, D.: Palindrome complexity. *Theoretical Computer Science* 292(1), 9–31 (2003)
2. Balková, L., Pelantová, E., Štěpán Starosta.: Proof of the Brlek-Reutenauer conjecture. *Theoretical Computer Science* 475, 120–125 (2013)
3. Berthé, V., Vuillon, L.: Tilings and rotations on the torus: a two-dimensional generalization of Sturmian sequences. *Discrete Mathematics* 223(1-3), 27–53 (2000)
4. Brlek, S., Hamel, S., Nivat, M., Reutenauer, C.: On the palindromic complexity of infinite words. *International Journal on Foundation of Computer Science* 15(2), 293–306 (2004)
5. Brlek, S., Reutenauer, C.: Complexity and palindromic defect of infinite words. *Theoretical Computer Science* 412(4-5), 493–497 (2011)
6. Crochemore, M., Iliopoulos, C.S., Kociumaka, T., Kubica, M., Radoszewski, J., Rytter, W., Tyczynski, W., Walen, T.: The maximum number of squares in a tree. In: *Combinatorial Pattern Matching - 23rd Annual Symposium, CPM 2012, Helsinki, Finland, July 3-5, 2012. Proceedings.* pp. 27–40 (2012)
7. de Luca, A.: Sturmian words: Structure, combinatorics, and their arithmetics. *Theoretical Computer Science* 183(1), 45–82 (1997)
8. Domenjoud, E., Provençal, X., Vuillon, L.: Palindromic language of thin discrete planes (To appear)
9. Domenjoud, E., Vuillon, L.: Geometric palindromic closure. *Uniform Distribution Theory* 7(2), 109–140 (2012)
10. Droubay, X., Justin, J., Pirillo, G.: Episturmian words and some constructions of de Luca and Rauzy. *Theoretical Computer Science* 255(1-2), 539–553 (2001)
11. Erdős, P., Turán, P.: On a problem of Sidon in additive number theory, and on some related problems. *Journal of the London Mathematical Society. Second Series* 16, 212–215 (1941)
12. Fraenkel, A.S., Simpson, J.: How many squares can a string contain? *J. Combin. Theory Ser. A* 82(1), 112–120 (1998)
13. Glen, A., Justin, J.: Episturmian words: a survey. *Theoretical Informatics and Applications. Informatique Théorique et Applications* 43(3), 403–442 (2009)
14. Gowers, T.: What are dense Sidon subsets of  $\{1, 2, \dots, n\}$  like? (2012), [gowers.wordpress.com/2012/07/13/what-are-dense-sidon-subsets-of-1-2-n-like/](http://gowers.wordpress.com/2012/07/13/what-are-dense-sidon-subsets-of-1-2-n-like/)
15. Hof, A., Knill, O., Simon, B.: Singular continuous spectrum for palindromic Schrödinger operators. *Comm. in Mathematical Physics* 174(1), 149–159 (1995)
16. Labbé, S., Reutenauer, C.: A  $d$ -dimensional extension of Christoffel words. *Discrete & Computational Geometry* (2015), preprint : <http://arxiv.org/abs/1404.4021>
17. Lothaire, M.: *Combinatorics on Words*. Cambridge University Press, Cambridge (1997)